The Annals of Statistics
2012, Vol. 40, No. 1, 73–103
DOI: 10.1214/11-AOS947

UPS DELIVERS OPTIMAL PHASE DIAGRAM IN HIGH-DIMENSIONAL VARIABLE SELECTION

BY PENGSHENG JI[1,2] AND JIASHUN JIN[1]

Cornell University and Carnegie Mellon University

Consider a linear model $Y = X\beta + z$, $z \sim N(0, I_n)$. Here, $X = X_{n,p}$, where both p and n are large, but $p > n$. We model the rows of X as i.i.d. samples from $N(0, \frac{1}{n}\Omega)$, where Ω is a $p \times p$ correlation matrix, which is unknown to us but is presumably sparse. The vector β is also unknown but has relatively few nonzero coordinates, and we are interested in identifying these nonzeros.

We propose the Univariate Penalization Screeing (UPS) for variable selection. This is a screen and clean method where we screen with univariate thresholding and clean with penalized MLE. It has two important properties: sure screening and separable after screening. These properties enable us to reduce the original regression problem to many small-size regression problems that can be fitted separately. The UPS is effective both in theory and in computation.

We measure the performance of a procedure by the Hamming distance, and use an asymptotic framework where $p \to \infty$ and other quantities (e.g., n, sparsity level and strength of signals) are linked to p by fixed parameters. We find that in many cases, the UPS achieves the optimal rate of convergence. Also, for many different Ω, there is a common three-phase diagram in the two-dimensional phase space quantifying the signal sparsity and signal strength. In the first phase, it is possible to recover all signals. In the second phase, it is possible to recover most of the signals, but not all of them. In the third phase, successful variable selection is impossible. UPS partitions the phase space in the same way that the optimal procedures do, and recovers most of the signals as long as successful variable selection is possible.

The lasso and the subset selection are well-known approaches to variable selection. However, somewhat surprisingly, there are regions in the phase space where neither of them is rate optimal, even in very simple settings, such as Ω is tridiagonal, and when the tuning parameter is ideally set.

Received May 2011; revised November 2011.
[1]Supported in part by NSF CAREER Award DMS-09-08613.
[2]Supported in part by NSF Grant DMS-08-05632.
AMS 2000 subject classifications. Primary 62J05, 62J07; secondary 62G20, 62C05.
Key words and phrases. Graph, Hamming distance, lasso, Stein's normal means, penalization methods, phase diagram, screen and clean, subset selection, variable selection.

1. Introduction. Consider the following sequence of regression problems:

$$Y^{(p)} = X^{(p)}\beta^{(p)} + z^{(p)}, \qquad z^{(p)} \sim N(0, I_n), \qquad n = n_p. \tag{1.1}$$

Here, $X^{(p)}$ is an $n_p \times p$ matrix, where both p and n_p are large, but $p > n_p$. The $p \times 1$ vector $\beta^{(p)}$ is unknown to us, but is sparse in the sense that it has s_p nonzeros where $s_p \ll p$. We are interested in variable selection: determining which components of $\beta^{(p)}$ are nonzero. For notational simplicity, we suppress the superscript $^{(p)}$ and subscript p whenever there is no confusion.

A well-known approach to variable selection is *subset selection*, also known as the L^0-penalization method (e.g., AIC [2], BIC [23] and RIC [13]). This approach selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \frac{(\lambda^{\mathrm{ss}})^2}{2}\|\beta\|_0, \tag{1.2}$$

where $\lambda^{\mathrm{ss}} > 0$ is a tuning parameter, and $\|\cdot\|_q$ denotes the L^q-norm. The approach has good properties, but the optimization problem (1.2) is known to be NP hard, which prohibits the use of the approach when p is large.

In the middle 1990s, Tibshirani [24] and Chen et al. [6] proposed a trail-breaking approach which is now known as the lasso or the basis pursuit. This approach selects variables by minimizing a similar functional, but $\|\beta\|_0$ is replaced by $\|\beta\|_1$.

$$\tfrac{1}{2}\|Y - X\beta\|_2^2 + \lambda^{\mathrm{lasso}}\|\beta\|_1. \tag{1.3}$$

A major advantage of the lasso is that (1.3) can be efficiently solved by the interior point method [6], even when p is relatively large. Additionally, in a series of papers (e.g., [9, 10]), it was shown that in the noiseless case (i.e., $z = 0$), the lasso solution is also the subset selection solution, provided that β is sufficiently sparse. For these reasons, the lasso procedure is passionately embraced by statisticians, engineers, biologists and many others.

With that being said, an obvious shortcoming of these methods is that the penalization term does not reflect the correlation structure in X, which prohibits the method from fully capturing the essence of the data (e.g., Zou [30]). However, this shortcoming is largely due to that these methods are *one-stage* procedures. This calls for a *two-stage* or *multi-stage* procedure.

1.1. *Screen and clean.* An idea introduced in the 1960s, screen and clean, has seen a revival recently [12, 27]. This is a two-stage method, where, at the first stage, we remove as many irrelevant variables as possible while keeping all relevant ones. At the second stage, we reinvestigate the surviving variables in hope of removing all false positives. The screening stage has the following advantages, some of which are elaborated in the literature:

- *Dimension reduction.* We remove many irrelevant variables, reducing the dimension from p to a much smaller number [12, 27].

- *Correlation complexity reduction.* A variable may be correlated to many other variables, but few of which will survive the screening; it is only correlated with a few other surviving variables.
- *Computation complexity reduction.* Under some conditions (e.g., Section 2), surviving variables can be grouped into many small units, each has a size $\leq K$, and correlation between units is weak. These units can be fitted separately, with computational cost $\leq$ # of units $\times 2^K$.

Despite the perceptive vision and philosophical importance in these works [12, 27], substantial vagueness remains: How to screen? How to clean? Is screen and clean really better than the lasso and the subset selection? This is where the Univariate Penalization Screening (UPS) comes in.

1.2. *UPS.* The UPS is a two-stage method which contains an U-step and a P-step. In the U-step, we screen with univariate thresholding [9] (also known as marginal regression [15] and sure screening [12]). Fix a threshold $t > 0$, and let x_j be the jth column of X. We remove the jth variable from the regression model if and only if $|(x_j, Y)| < t$. The set of surviving indices is then $\mathcal{U}_p(t) = \mathcal{U}_p(t; Y, X) = \{j : |(x_j, Y)| \geq t, 1 \leq j \leq p\}$.

Despite its simplicity, the U-step can be effective in many situations. The key insight is that $\mathcal{U}_p(t)$ has the following important properties:

- *Sure Screening* (*SS*). With overwhelming probability, $\mathcal{U}_p(t)$ includes all but a negligible proportion of the signals (i.e., nonzero coordinates of β). The terminology is slightly different from that in [12].
- *Separable After Screening* (*SAS*). Define a graph where $\{1, 2, \ldots, p\}$ is the set of nodes, and nodes j and k are connected if and only if $|(x_j, x_k)|$ is large (i.e., columns j and k are "significantly" correlated). The SAS property refers to as that with overwhelming probability, $\mathcal{U}_p(t)$ splits into many disconnected small-size components [a component is a maximal connected subgraph of $\mathcal{U}_p(t)$].

We now explain how these properties pave the way for the P-step. Let $\mathcal{I}_0 = \{i_1, \ldots, i_K\}$ and $\mathcal{J}_0 = \{j_1, \ldots, j_L\}$ be two subsets of $\{1, 2, \ldots, p\}$, $1 \leq K$, $L \leq p$. We have the following definition.

DEFINITION 1.1. For any $p \times 1$ vector Y, $Y^{\mathcal{I}_0}$ denotes the $K \times 1$ vector such that $Y^{\mathcal{I}_0}(k) = Y_{i_k}$, $1 \leq k \leq K$. For any $p \times p$ matrix Ω, $\Omega^{\mathcal{I}_0, \mathcal{J}_0}$ denotes the $K \times L$ matrix such that $\Omega^{\mathcal{I}_0, \mathcal{J}_0}(k, \ell) = \Omega(i_k, j_\ell)$, $1 \leq k \leq K, 1 \leq \ell \leq L$.

Note that the regression model is closely related to the model $X'Y = X'X\beta + X'z$. Restricting the attention to $\mathcal{U} = \mathcal{U}_p(t)$, we have

$$(X'Y)^{\mathcal{U}} = (X'X\beta)^{\mathcal{U}} + (X'z)^{\mathcal{U}} = (X'X)^{\mathcal{U},\mathcal{V}}\beta + (X'z)^{\mathcal{U}},$$

where $\mathcal{V} = \{1, 2, \ldots, p\}$. Three key observations are the following: (a) since $z \sim N(0, I_n)$, $(X'z)^{\mathcal{U}} \sim N(0, (X'X)^{\mathcal{U},\mathcal{U}})$, (b) by the sure screening property,

$(X'X)^{\mathcal{U},\mathcal{V}}\beta \approx (X'X)^{\mathcal{U},\mathcal{U}}\beta^{\mathcal{U}}$ and (c) by the SAS property, $(X'X)^{\mathcal{U},\mathcal{U}}$ approximately equals a block diagonal matrix, where each block corresponds to a maximal connected subgraph contained in $\mathcal{U}_p(t)$. As a result, the original regression problem reduces to many small-size regression problems that can be solved separately, each at a modest computational cost.

In detail, fix two parameters λ^{ups} and u^{ups}. Let $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \subset \mathcal{U}_p(t)$ be a component, and let μ be a $K \times 1$ vector the coordinates of which are either 0 or u^{ups}. Write $A = (X'X)^{\mathcal{I}_0,\mathcal{I}_0}$ for short. Let $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}}, p)$ be the minimizer of the functional

$$\tfrac{1}{2}((X'Y)^{\mathcal{I}_0} - A\mu)'A^{-1}((X'Y)^{\mathcal{I}_0} - A\mu) + \tfrac{1}{2}(\lambda^{\mathrm{ups}})^2\|\mu\|_0. \tag{1.4}$$

Combining all such estimates across different components of $\mathcal{U}_p(t)$ gives the UPS estimator, denoted by $\hat{\beta}^{\mathrm{ups}} = \hat{\beta}^{\mathrm{ups}}(Y, X; t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}}, p)$,

$$\hat{\beta}_j^{\mathrm{ups}} = \begin{cases} (\hat{\mu}(\mathcal{I}_0))_k, & \text{if } j = i_k \in \mathcal{I}_0 \text{ for some } \mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \subset \mathcal{U}_p(t), \\ 0, & \text{if } j \notin \mathcal{U}_p(t_p). \end{cases}$$

The UPS uses three tuning parameters $(t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}})$. In many cases, the performance of the UPS is relatively insensitive to the choice of t, as long as it falls in a certain range. The parameter λ^{ups} has a similar role to those of the lasso and the subset selection, but there is a major difference: the former can be conveniently estimated using the data, whereas how to set the latter remains an open problem. See Section 2 for more discussion.

We are now ready to answer the questions raised in the end of Section 1.1: UPS indeed has advantages over the lasso and the subset selection. In Sections 1.3–1.7, we establish a theoretic framework and investigate these procedures closely. The main finding is the following: for a wide range of design matrices X, the Hamming distance of the UPS achieves the optimal rate of convergence. In contrast, the lasso and the subset selection may be rate nonoptimal, even for very simple design matrices.

1.3. *Sparse signal model and universal lower bound.* We model β by

$$\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)\nu_0 + \varepsilon\pi, \qquad 0 < \varepsilon < 1, 1 \le j \le p, \tag{1.5}$$

where ν_0 is the point mass at 0, and π is a distribution that has no mass at 0. We use p as the driving asymptotic parameter and allow (ε, π) to depend on p. Fix $0 < \vartheta < 1$ and recall that s_p is the number of signals. We calibrate

$$\varepsilon = \varepsilon_p = p^{-\vartheta} \qquad \text{so that } s_p \sim p\varepsilon_p = p^{1-\vartheta}. \tag{1.6}$$

For any variable selection procedure $\hat{\beta} = \hat{\beta}(Y|X)$, we measure the loss by the Hamming distance

$$h_p(\hat{\beta}, \beta|X) = h_p(\hat{\beta}, \beta; \varepsilon_p, \pi_p, n_p|X) = E_{\varepsilon_p,\pi_p}\left[\sum_{j=1}^{p} 1(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j))\right],$$

where $\operatorname{sgn}(0) = 0$. In the context of variable selection, the Hamming distance is a natural choice for loss function. While the focus of this paper is on selection error where we use L_0-loss, the idea can be extended to the estimation setting where we use L_q-loss ($0 < q < \infty$), but we have to perform an additional step of least square fitting after the selection.

Somewhat surprisingly, there is a lower bound for the Hamming distance that holds for all sample size n and design matrix X (and so "universal lower bound"). The following notation is frequently used in this paper.

DEFINITION 1.2. $L_p > 0$ is a multi-$\log(p)$ term which may change from occurrence to occurrence, such that for any fixed $\delta > 0$, $\lim_{p\to\infty} L_p \cdot p^{\delta} = \infty$ and $\lim_{p\to\infty} L_p p^{-\delta} = 0$.

Now, fixing $r > 0$, we introduce

$$\tau_p = \tau_p(r) = \sqrt{2r\log p} \tag{1.7}$$

and $\lambda_p = \lambda_p(\varepsilon_p, \tau_p) = \frac{1}{\tau_p}[\log(\frac{1-\varepsilon_p}{\varepsilon_p}) + \frac{\tau_p^2}{2}]$. Let $\bar{\Phi} = 1 - \Phi$ be the survival function of $N(0,1)$. The following theorem is proved in [18].

THEOREM 1.1 (Lower bound). *Fix $\vartheta \in (0,1)$, $r > 0$ and a sufficiently large p. Let ε_p, s_p and τ_p be as in (1.6) and (1.7), and suppose the support of π_p is contained in $[-\tau_p, 0) \cup (0, \tau_p]$. For any fixed n and matrix $X = X^{(p)}$ such that $X'X$ has unit diagonals, $h_p(\hat{\beta}, \beta|X) \geq s_p \cdot [(1-\varepsilon_p)\bar{\Phi}(\lambda_p)/\varepsilon_p + \Phi(\tau_p - \lambda_p)]$.*

Note that as $p \to \infty$,

$$\frac{1-\varepsilon_p}{\varepsilon_p}\bar{\Phi}(\lambda_p) + \Phi(\tau_p - \lambda_p) \geq \begin{cases} L_p \cdot p^{-(r-\vartheta)^2/(4r)}, & r > \vartheta, \\ (1+o(1)), & r < \vartheta. \end{cases} \tag{1.8}$$

It may seem counterintuitive that the lower bound does not depend on n, but this is due to the way we normalize X. In the case of orthogonal design [i.e., coordinates of X and i.i.d. from $N(0, 1/n)$], the lower bound can be achieved by either the lasso or marginal regression [15]. Therefore, the orthogonal design is among the best in terms of the error rate.

Theorem 1.1 says that if we have $p^{1-\vartheta}$ signals, and the maximal signal strength is slightly smaller than $\sqrt{2\vartheta\log(p)}$, then the Hamming distance of any procedure cannot be substantially smaller than s_p, and so successful variable selection is impossible. In the sections below, we focus on the case where the signal strength is larger than $\sqrt{2\vartheta\log(p)}$, so that successful variable selection is possible.

The universality of the lower bound hints it may not be tight for nonorthogonal X. Fortunately, it turns out that in many interesting cases, the lower bound is tight. To facilitate the analysis, we invoke the random design model.

1.4. *Random design, connection to Stein's normal means model.* Write $X = (x_1, x_2, \ldots, x_p) = (X_1, X_2, \ldots, X_n)'$. We model X_i as i.i.d. samples from a p-variate zero-mean Gaussian distribution,

$$X_i \overset{\text{i.i.d.}}{\sim} N\Big(0, \frac{1}{n}\Omega\Big). \tag{1.9}$$

The $p \times p$ matrix $\Omega = \Omega^{(p)}$ is unknown, but for simplicity we assume it has unit diagonals. The normalizing constant $1/n$ is chosen so that the diagonals of the Gram matrix $X'X$ are approximately 1. Fixing $\theta \in (1 - \vartheta, 1)$, we let

$$n = n_p = p^{\theta}. \tag{1.10}$$

Note that $s_p \ll n_p \ll p$ as $p \to \infty$. For successful variable selection, it is almost necessary to have $s_p \ll n_p$ [9]. Also, denoting the distribution of X by $F = F_p$, note that for any variable selection procedure, the *overall Hamming distance* is $\mathrm{Hamm}_p(\hat{\beta}, \beta) = E_F[h_p(\hat{\beta}|X)]$.

Model (1.9) is called the *random design model* which may be found in the following application areas:

- *Compressive sensing.* We are interested in a p-dimensional sparse vector β. We measure n general linear combinations of β and then reconstruct it. For $1 \le i \le n$, choose a $p \times 1$ coefficient vector X_i, and observe $Y_i = X_i'\beta + z_i$, where $z_i \sim N(0, \sigma^2)$ is noise. For computational and storage concerns, one usually chooses X_i's as simple as possible. Popular choices of X_i include Gaussian design, Bernoulli design, circulant design, etc. [3, 9]. Model (1.9) belongs to Gaussian design.
- *Privacy-preserving data mining.* The vector β may contain some confidential information (e.g., HIV-diagnosis results of a community) that we must protect. While we cannot release the whole vector, we must allow data mining to some extent, because, for example, the study is of public interest and is supported by federal funding. To compromise, we allow queries as follows. For each query, the database randomly generates a $p \times 1$ vector X_i, and releases both X_i and $Y_i = X_i'\beta + z_i$ to the querier, where $z_i \sim N(0, \sigma^2)$ is a noise term. For privacy concerns, the number of allowed queries is much smaller than p. Popular choices of X_i include Gaussian design and Bernoulli design [8].

Random design model is closely related to Stein's normal means model $W \sim N(\beta, \Sigma)$, where $\Sigma = \Omega^{-1}$. To see the point, recall that model (1.1) is closely related to the model $X'Y = X'X\beta + X'z$. Since the rows of X are i.i.d. samples from $N(0, \frac{1}{n}\Omega)$ and $s_p \ll n_p \ll p$, we expect to see that $X'X\beta \approx \Omega\beta$ and $X'z \approx N(0, \Omega)$, and so that $X'Y \approx N(\Omega\beta, \Omega)$. Therefore, Stein's normal means model can be viewed as an idealized version of the random design model. This suggests that solving the variable selection problem opens doors for solving Stein's normal means problem, and vice versa.

1.5. *Optimality of the UPS.* The main results of this paper are Theorems 2.1 and 2.2 in Section 2. To state such results, we need relatively long preparations. Therefore, we sketch these results below, but leave the formal statements to later. In models (1.1), (1.5) and (1.9), let (s_p, τ_p, n_p) be as in (1.6), (1.7) and (1.10). Suppose:

- Each row of Ω satisfies a certain summability condition, so it has relatively few large coordinates.
- The support of π_p is contained in $[\tau_p, (1+\eta)\tau_p]$, where $\tau_p = \sqrt{2r\log(p)}$, and η is a constant to be defined later. We suppose $r > \vartheta$, so that successful variable selection is possible; see Theorem 1.1.
- Either all coordinates of Ω are positive, or that $r/\vartheta \le 3 + 2\sqrt{2}$ (so that we won't have too many "signal cancellations" [27]).

Fix $0 < q \le (\vartheta + r)^2/(4r)$, and set the tuning parameters $(t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}})$ by

$$t_p^* = t_p^*(q) = \sqrt{2q\log p}, \qquad \lambda^{\mathrm{ups}} = \lambda_p^{\mathrm{ups}} = \sqrt{2\vartheta\log(p)}, \qquad u^{\mathrm{ups}} = u_p^{\mathrm{ups}} = \tau_p.$$

The main result is that, as $p \to \infty$, the ratio between the Hamming error of the UPS and s_p is no grater than $L_p p^{-(\vartheta - r)^2/(4r)}$. Comparing this with Theorem 1.1 gives that the lower bound is tight, and the UPS is rate optimal.

1.6. *Phase diagram for high-dimensional variable selection.* The above results reveal a watershed phenomenon as follows. Suppose we have roughly $s_p = p^{1-\vartheta}$ signals. If the maximal signal strength is slightly smaller than $\sqrt{2\vartheta\log p}$, then the Hamming distance of any procedure cannot be substantially smaller than s_p, hence successful variable selection is impossible. If the minimal signal strength is slightly larger than $\sqrt{2\vartheta\log p}$, then there exist procedures (UPS is one of them) whose Hamming distances are substantially smaller than s_p, and they manage to recover most signals.

The phenomenon is best described in the special case where $\pi_p = \nu_{\tau_p}$ is the point mass at τ_p, with $\tau_p = \sqrt{2r\log p}$ as in (1.7). If we call the two-dimensional domain $\{(\vartheta, r) : 0 < \vartheta < 1, r > 0\}$ the *phase space*, then the theorems say that the phase space is partitioned into three regions:

- *Region of no recovery* $(0 < \vartheta < 1,\ 0 < r < \vartheta)$. In this region, the Hamming distance of any procedure $\gtrsim s_p$, and successful variable selection is impossible.
- *Region of almost full recovery* $[0 < \vartheta < 1,\ \vartheta < r < (1 + \sqrt{1-\vartheta})^2]$. In this region, there are procedures (e.g., UPS) whose Hamming errors are much larger than 1, but are also much smaller than s_p. In this region, it is possible to recover most of the signals, but not all of them.
- *Region of exact recovery* $[0 < \vartheta < 1,\ r > (1 + \sqrt{1-\vartheta})^2]$. In this region, there are procedures (e.g., UPS) that recover all signals with probability ≈ 1.

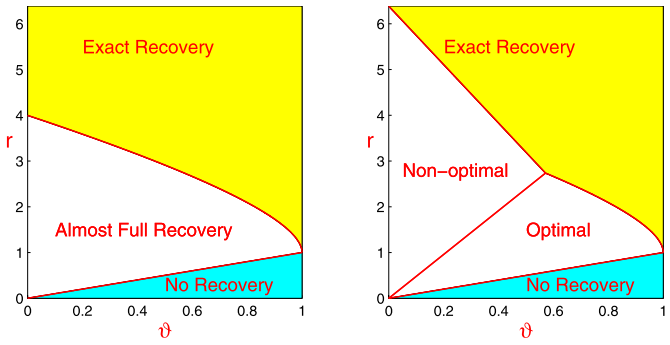

FIG. 1. *Left: phase diagram. In the yellow region, the UPS recovers all signals with high probability. In the white region, it is possible (i.e., UPS) to recover almost all signals, but impossible to recover all of them. In the cyan region, successful variable selection is impossible. Right: partition of the phase space by the lasso for the tridiagonal model (1.11)–(1.12) ($a = 0.4$). The lasso is rate nonoptimal in the nonoptimal region. The region of exact recovery by the lasso is substantially smaller than that displayed on the left.*

See Figure 1 (left panel) for these regions. Note that the partitions are the same for many choices of Ω. Because of the partition of the phases, we call this the phase diagram. The UPS is optimal in the sense that it partitions the phase space in exactly the same way as do the optimal procedures.

The phase diagram provides a benchmark for variable selection. The lasso would be optimal if it partitions the phase space in the same way as in the left panel of Figure 1. Unfortunately, this is not the case, even for very simple Ω. Below we investigate the case where $X'X$ is a tridiagonal matrix, and identify precisely the regions where the lasso is rate optimal and where it is rate nonoptimal. More surprisingly, there is a region in the phase space where the subset selection is also rate nonoptimal.

1.7. *Nonoptimal region for the lasso.* In Sections 1.7 and 1.8, we temporarily leave the random design model and consider Stein's normal means model, which is an idealized version of the former. Using an idealized version is mainly for mathematical convenience, but the gained insight is valid in much broader settings: if a procedure is nonoptimal in simple cases, we should not expect them to be optimal in more complicated cases.

In this spirit, we consider Stein's normal means model

$$\tilde{Y} \equiv X'Y \sim N(\Omega\beta, \Omega), \tag{1.11}$$

where β is as in (1.5) with $\tau_p = \nu_{\pi_p}$ and $\pi_p = \sqrt{2r\log(p)}$. To further simplify the study, we fix $a \in (0, 1/2)$ and take Ω as the tridiagonal matrix $T(a)$:

$$T(a)(i,j) = 1\{i = j\} + a \cdot 1\{|i - j| = 1\}, \qquad 1 \le i, j \le p. \tag{1.12}$$

Note that in this case the UPS partitions the phase space optimally.

We now discuss the phase diagram of the lasso. The region $\{(\vartheta, r) : 0 < \vartheta < 1, r > \vartheta\}$ is partitioned into three regions as follows (see Figure 1):

- *Nonoptimal region*: $0 < \vartheta < 2a(1+a)^{-1}$ and $\frac{1}{a}(1+\sqrt{1-a^2})\vartheta < r < (1+\sqrt{\frac{1+a}{1-a}})^2(1-\vartheta)$. In this region, the lasso is rate nonoptimal [i.e., the Hamming distance is $L_p \cdot p^c$ with constant $c > 1 - (\vartheta + r)^2/(4r)$], even when the tuning parameter is set ideally.
- *Optimal region*: $0 < \vartheta < 1$ and $\vartheta < r < \frac{1}{a}(1+\sqrt{1-a^2})\vartheta$ and $r < (1+\sqrt{1-\vartheta})^2$. In this region, if additionally $a \geq 1/3$, then the lasso may be rate optimal if the tuning parameter is set ideally. The discussion on the case $0 < a < 1/3$ is tedious so we skip it.
- *Region of exact recovery*: $0 < \vartheta < 1$ and $r > (1+\sqrt{1-\vartheta})^2$ and $r > (1+\sqrt{\frac{1+a}{1-a}})^2(1-\vartheta)$. In this region, if the tuning parameter is set ideally, the lasso may yield exact recovery with high probability. Region of exactly recovery by the lasso is substantially smaller than that of the UPS. There is a sub-region in the phase space where the UPS yields exact recovery, but the lasso could not even when the tuning parameter is set ideally.

For discussions in the case where Ω is the identity matrix, compare [15, 25]. The above results are proved in Theorem 4.1, where we derive a lower bound for the Hamming errors by the lasso. In [17], we show that the lower bound is tight for properly large ϑ, but is not when ϑ is small. It is, however, tight for all $\vartheta \in (0,1)$ if we replace model (1.5) by a closely related model, namely (2.2) and (2.3) in [16]. For these reasons, the nonoptimal region of the lasso may be larger than that illustrated in Figure 1. The discussion on the exact optimal rate of convergence for the lasso is tedious and we skip it.

Why is the lasso nonoptimal? To gain insight, we introduce the term of *fake signal*, a noise coordinate that may look like a signal due to correlation.

DEFINITION 1.3. We say that $\tilde{Y}_j$ is a signal if $\beta_j \neq 0$, is a fake signal if $(\Omega\beta)_j \neq 0$ and $\beta_j = 0$, and is a (pure) noise if $\beta_j = (\Omega\beta)_j = 0$.

With the tuning parameter set ideally, the lasso is able to distinguish signals from pure noise, but it does not filter out fake signals efficiently. In the optimal region of the lasso, the number of falsely kept fake signals is much smaller than the optimal rate, so it is negligible; in the nonoptimal region, the number becomes much larger than the optimal rate, and so is nonnegligible. This suggests that when $X'X$ moves away from the tridiagonal case, the partitions of the regions by the lasso may change, but the nonoptimal region of the lasso continues to exist in rather general situations.

The nonoptimality of the lasso is largely due to the fact that it is a one-stage method. An interesting question is whether UPS continues to work well if we replace the univariate thresholding by the lasso in the screening stage. The disadvantage of this proposal is that, compared to the univariate

thresholding, the lasso is both slower in computation and harder to analyze in theory. Still, one would hope the lasso could perform well in screening.

With that being said, we note that the implementation of the lasso only needs minimal assumption on the model, which makes it very attractive, especially in complicated situations. In comparison, we need both signal sparsity and graph sparsity to implement the UPS, and how to extend it to more general settings remains unknown. The exploration along this line is continued in our forthcoming manuscripts [11, 19, 20]; see details therein.

1.8. *Nonoptimal region for the subset selection.* The discussion on the subset selection is similar to that for the lasso so we keep it brief. Introduce $v_1(a) = \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}$ and $v_2(a) = 2\sqrt{1-a^2}-1$. Similarly, the phase space partitions into three regions as follows:

- *Nonoptimal region*: $0 < \vartheta < \frac{4v_1(a)}{(v_1(a)+1)^2}$ and $v_1(a)\vartheta < r < [\frac{1}{v_2(a)}(\sqrt{1-2\vartheta} + \sqrt{1-2\vartheta+\vartheta v_2(a)})]^2$.
- *Optimal region*: $0 < \vartheta < 1$ and $\vartheta < r < v_1(a)\vartheta$ and $r < (1+\sqrt{1-\vartheta})^2$.
- *Exact recovery region*: $0 < \vartheta < 1$, $r > (1+\sqrt{1-\vartheta})^2$ and $r > [\frac{1}{v_2(a)}(\sqrt{1-2\vartheta}+ \sqrt{1-2\vartheta+\vartheta v_2(a)})]^2$.

See Theorem 4.2 for proofs and Figure 2 for illustration. Similar to the remarks in Section 1.7, the region of exact recovery and the optimal region of the subset selection may be smaller than those illustrated in Figure 2.

The reason why the subset selection is nonoptimal is almost the *opposite* to that of the lasso: the lasso is nonoptimal for it is too loose on fake signals, but the subset selection is nonoptimal for it is too harsh on signal clusters

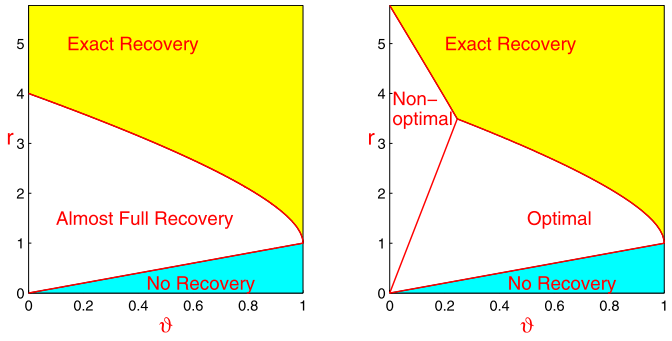

FIG. 2. *Left: a re-display of the left panel of Figure 1. Right: partition of the phase space by the subset selection in the tridiagonal model (1.11)–(1.12) ($a = 0.4$). The subset selection is not rate optimal in the nonoptimal region. The exact recovery region by the subset selection is substantially smaller than that of the optimal procedure, displayed on the left.*

(pairs/triplets, etc.). With the tuning parameter set ideally, the subset selection is effective in filtering out fake signals, but it also tends to kill one or more signals when the true signals appear in clusters. These falsely killed signals account for the nonoptimality. See Section 4.2 for details.

1.9. *Connection to recent literature.* This work is related to recent literature on oracle property [22, 30], but is different in important ways. A procedure has the oracle property if it yields exact recovery. However, exact recovery is rarely seen in applications, especially when $p \gg n$. In many applications (e.g., genomics), a large p usually means that signals are sparse or rare, and a small n usually means signals are weak. For rare and weak signals, exact recovery is usually impossible. Therefore, it is both scientifically more relevant and technically more challenging to compare error rates of different procedures than to investigate when they satisfy the oracle property.

The work is also related to [5, 28] on asymptotic minimaxity, where the lasso was shown to be asymptotic rate optimal in the worst-case scenario. While their results seem to contradict with those in this paper, the difference can be easily reconciled. In the minimax approach, the asymptotic least favorable distribution of β is given by $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p \nu_{\tau_p}$, where $\varepsilon_p = p^{-\vartheta}$, $\tau_p = \sqrt{2r\log p}$ and notably $\vartheta = r$, which corresponds the boundary line of the region of no recovery in the phase space (e.g., [28], pages 18 and 19, [1], Section 3). This suggests that the minimax approach has limitations: it reduces the analysis to the worst-case scenario, but the worst-case scenario may be outside the range of interest. In our approach, we let (ϑ, r) range freely, and evaluate a procedure based on how it partitions the phase space. Our approach has a similar spirit to that in [10].

The work is also related to the adaptive lasso [30]. The adaptive lasso is similar to the lasso, but the L^1-penalty $\lambda^{\text{lasso}}\|\beta\|_1$ is replaced by the weighted L^1-penalty $\sum_{j=1}^p w_j|\beta_j|$, where $w = (w_1, \ldots, w_p)'$ is the weight vector. Philosophically, we can view the adaptive lasso as a screen and clean method. Still, the proposed approach is different from the adaptive lasso in important ways. First, Zou [30] suggested weight choices by the least squares estimate, which is only feasible when p is small. In fact, when $p \gg n$, our results suggest that feasible weights should be very sparse, while the weights suggested by the least squares estimates are usually dense. Second, for the surviving indices, we first partition them into many disjoint units of small sizes, and then fit them individually. The adaptive lasso fits all surviving variables together, which is computationally more expensive. Last, we use penalized MLE in the clean step while the adaptive lasso uses L^1-penalty. As pointed out before, the L^1-penalty in the clean step is too loose on fake signals, which prohibits the procedure from being rate optimal.

The work is also related to other multi-stage methods, for example, the threshold lasso [29] or the LOL [21]. These methods first use the lasso and

the OLS for variable selection, respectively, followed by an additional thresholding step. However, by an argument similar to that in Sections 1.7 and 1.8, it is not hard to see that these procedures do not partition the phase diagram optimally.

1.10. *Contents.* In summary, we propose the UPS as a two-stage method for variable selection. We use Univariate thresholding in the screening step for its exceptional convenience in computation, and we use penalized MLE in the cleaning step because it is the only procedure we know so far that yields the optimal rate of convergence. On the other hand, the lasso and even the subset selection do not partition the phase space optimally.

The remaining sections are organized as follows. Section 2 discusses the UPS procedure and the upper bound for the rate of convergence. The section also addresses how to estimate the tuning parameters of the UPS and the convergence rate of the resultant plug-in procedure. Section 3 discusses a refinement of the UPS for moderately large p. Section 4 discusses the behavior of the lasso and the subset selection. Section 5 discusses numerical results where we compare the UPS with the lasso (the subset selection is computationally infeasible for large p so is not included for comparison). Due to limited space, we do not include proofs in this paper. The proofs can be found in the supplementary material for the paper [18].

Below is some notation we use in this paper. Fix $0 < q < \infty$. For a $p \times 1$ vector x, $\|x\|_q$ denotes the L^q-norm of x, and we omit the subscript when $q = 2$. For a $p \times p$ matrix M, $\|M\|_q$ denotes the matrix L^q-norm, and $\|M\|$ denotes the spectral norm.

2. UPS and upper bound for the Hamming distance. In this section, we establish the upper bound for the Hamming distance and show that the UPS is rate optimal. We begin by discussing necessary notation. We then discuss the U-step and its sure screening and SAS properties. Next, we show how the regression problem reduces to many separate small-size regression problems and explain the rationale of using the penalized MLE in the P-step. We conclude the section by the rate optimality of the UPS, where the tuning parameters are either set ideally or estimated.

Since different parts of our model are introduced separately in different subsections, we summarize them as follows. The model we consider is

$$Y = X\beta + z, \qquad z \sim N(0, I_n), \tag{2.1}$$

where

$$\begin{aligned} &X_i \overset{\text{i.i.d.}}{\sim} N\Big(0, \frac{1}{n}\Omega\Big), \\ &\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p \pi_p, \qquad 1 \le i \le n, 1 \le j \le p. \end{aligned} \tag{2.2}$$

Fixing $\theta > 0$, $\vartheta > 0$, and $r > 0$, we calibrate

$$\varepsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}, \qquad n_p = p^{\theta}, \tag{2.3}$$

assuming that

$$\theta < (1-\vartheta). \tag{2.4}$$

Recall that the optimal rate of convergence is $L_p p^{1-(\vartheta+r)^2/(4r)}$. In this section, we focus on the case where the exponent $1-(\vartheta+r)^2/(4r)$ falls between 0 and $(1-\vartheta)$, or equivalently,

$$\vartheta < r < (1+\sqrt{1-\vartheta})^2. \tag{2.5}$$

In the phase space, this corresponds to the region of almost full recovery. The case $r < \vartheta$ corresponds to the region of no recovery and is studied in Theorem 1.1. The case $r > (1+\sqrt{1-\vartheta})^2$ corresponds to the region of exact recovery. The discussion in this case is similar but is much easier, so we omit it.

Next, fixing $A > 0$ and $\gamma \in (0,1)$, introduce

$$\mathcal{M}_p(\gamma, A) = \left\{ \Omega : p \times p \text{ correlation matrix}, \sum_{j=1}^{p} |\Omega(i,j)|^{\gamma} \le A, \forall 1 \le i \le p \right\}.$$

For any Ω, let $U = U(\Omega)$ be the $p \times p$ matrix satisfying $U(i,j) = \Omega(i,j)1\{i < j\}$, and let $d(\Omega) = \max\{\|U(\Omega)\|_1, \|U(\Omega)\|_\infty\}$. Fixing $\omega_0 \in (0, 1/2)$, introduce $\mathcal{M}_p^*(\omega_0, \gamma, A) = \{\Omega \in \mathcal{M}_p(\gamma, A) : d(\Omega) \le \omega_0\}$, and a subset of $\mathcal{M}_p^*(\omega_0, \gamma, A)$,

$$\mathcal{M}_p^+(\omega_0, \gamma, A) = \{\Omega \in \mathcal{M}_p^*(\omega_0, \gamma, A) : \Omega(i,j) \ge 0 \text{ for all } 1 \le i, j \le p\}.$$

For any $\Omega \in \mathcal{M}_p^*(\omega_0, \gamma, A)$, the eigenvalues are contained in $(1-2\omega_0, 1+2\omega_0)$, so Ω is positive definite (when $\omega_0 > 1/2$, Ω may not be positive definite).

Last, introduce a constant $\eta = \eta(\vartheta, r, \omega_0)$ by

$$\eta = \frac{\sqrt{\vartheta r}}{(\vartheta + r)\sqrt{1+2\omega_0}} \min\left\{ \frac{2\vartheta}{r}, 1 - \frac{\vartheta}{r}, \sqrt{2(1-\omega_0)} - 1 + \frac{\vartheta}{r} \right\}. \tag{2.6}$$

We suppose the support of signal distribution π_p is contained in

$$[\tau_p, (1+\eta)\tau_p], \tag{2.7}$$

where $\tau_p = \sqrt{2r\log(p)}$ as in (1.7). This assumption is only needed for proving the main lemma of the P-step (Lemma A.5, [18]) and can be relaxed for proving other lemmas. Also, we assume the signals are one-sided mainly for simplicity. The results can be extended to the case with two-sided signals.

We now discuss the U-step. As mentioned before, the benefits of the U-step are threefold: dimension reduction, correlation complexity reduction, and computation cost reduction. The U-step is able to achieve these goals simultaneously because it satisfies the sure screening property and the SAS property, which we now discuss separately.

2.1. *The sure screening property of the U-step.* Recall that in the U-step, we remove the jth variable if and only if $|(x_j, Y)| < t$ for some threshold $t > 0$. For simplicity, we make a slight change and remove the jth variable if and only if $(x_j, Y) < t$. When the signals are one-sided, the change makes negligible difference. Fixing a constant $q \in (0, (\vartheta + r)^2/(4r))$, we set the threshold t in the U-step

$$t_p^* = t_p^*(q) = \sqrt{2q \log(p)}. \tag{2.8}$$

LEMMA 2.1 (Sure screening). *In model (2.1)–(2.2), suppose (2.3)–(2.7) hold, and t_p^* is as in (2.8). For sufficiently large p, if $\Omega^{(p)} \in \mathcal{M}_p^+(\omega_0, \gamma, A)$, then as $p \to \infty$, $\sum_{j=1}^p P(x_j' Y < t_p^*, \beta_j \neq 0) \leq L_p p^{1-(\vartheta+r)^2/(4r)}$. The claim remains true if alternatively $\Omega^{(p)} \in \mathcal{M}_p^*(\omega_0, \gamma, A)$, but $r/\vartheta \leq 3 + 2\sqrt{2}$.*

This says that the Hamming errors we make in the U-step are not substantially larger than the optimal rate of convergence, and thus negligible.

2.2. *The SAS property of the U-step.* We need some terminology in graph theory (e.g., [7]). A graph $G = (V, E)$ consists of two finite sets V and E, where V is the set of *nodes*, and E is the set of *edges*. A *component* $\mathcal{I}_0$ of V is a maximal connected subgraph, denoted by $\mathcal{I}_0 \lhd V$. For any node $v \in V$, there is a unique component $\mathcal{I}_0$ such that $v \in \mathcal{I}_0 \lhd V$.

Fix a $p \times p$ symmetric matrix Ω_0 which is presumably sparse. If we let $V_0 = \{1, 2, \ldots, p\}$ and say nodes i and j are *linked* if and only if $\Omega_0(i, j) \neq 0$, then we have a graph $G = (V_0, \Omega_0)$. Fix $t > 0$. Recall that $\mathcal{U}_p(t)$ is the set of surviving indices in the U-step

$$\mathcal{U}_p(t) = \mathcal{U}_p(t, Y, X) = \{j : (x_j, Y) \geq t, 1 \leq j \leq p\}. \tag{2.9}$$

Note that the induced graph $(\mathcal{U}_p(t), \Omega_0)$ splits into many components.

DEFINITION 2.1. Fix an integer $K \geq 1$. We say that $\mathcal{U}_p(t)$ has the separable after screening (SAS) property with respect to (V_0, Ω_0, K) if each component of the graph $(\mathcal{U}_p(t), \Omega_0)$ has no more than K nodes.

Note that if $\mathcal{U}_p(t)$ has the SAS property with respect to (V_0, Ω_0, K). Then for all $s > t$, $\mathcal{U}_p(s)$ also has the SAS property with respect to (V_0, Ω_0, K).

Return to model (2.1)–(2.2). We hope to relate the regression setting to a graph (V_0, Ω_0), and use it to spell out the SAS property. Toward this end, we set $V_0 = \{1, 2, \ldots, p\}$. As for Ω_0, a natural choice is the matrix Ω in (2.2). However, the SAS property makes more sense if Ω_0 is sparse and known, while Ω is neither. In light of this, we take Ω_0 to be a regularized empirical covariance matrix.

In detail, let $\hat{\Omega} = X'X$ be the empirical covariance matrix. Recall that $X = (X_1, X_2, \ldots, X_n)'$ and $X_i \sim N(0, \frac{1}{n}\Omega)$. It is known [4] that there is a constant

$C>0$ such that with probability $1-o(1/p^2)$, for all $1\le i,j\le p$,

$$|\hat{\Omega}(i,j)-\Omega(i,j)|\le C\sqrt{\log(p)}/\sqrt{n}. \tag{2.10}$$

For large p, $\hat{\Omega}$ is a noisy estimate for Ω, so we regularize it by

$$\Omega^*(i,j)=\hat{\Omega}(i,j)1_{\{|\hat{\Omega}(i,j)|\ge \log^{-1}(p)\}}. \tag{2.11}$$

The threshold $\log^{-1}(p)$ is chosen mainly for simplicity and can be replaced by $\log^{-a}(p)$, where $a>0$ is a constant. The following lemma is a direct result of (2.10); we omit the proof.

LEMMA 2.2. *Fix $A>0$, $\gamma\in(0,1)$ and $\omega_0\in(0,1/2)$. As $p\to\infty$, for any $\Omega\in\mathcal{M}_p^*(\omega_0,\gamma,A)$, with probability of $1-o(1/p^2)$, each row of Ω^* has no more than $2\log(p)$ nonzero coordinates, and $\|\Omega^*-\Omega\|_\infty\le C(\log(p))^{-(1-\gamma)}$.*

Taking $\Omega_0=\Omega^*$, we form a graph (V_0,Ω^*). The following lemma is proved in [18], which says that, except for a negligible probability, $\mathcal{U}_p(t_p^*)$ has the SAS property.

LEMMA 2.3 (SAS). *Consider model (2.1)–(2.2) where (2.3)–(2.7) hold. Set t_p^* as (2.8). As $p\to\infty$, there is a constant K such that with probability $1-L_pp^{-(\vartheta+r)^2/(4r)}$, $\mathcal{U}_p(t_p^*)$ has the SAS property with respect to (V_0,Ω^*,K).*

2.3. *Reduction to many small-size regression problems.* Together, the sure screening property and the SAS property make sure that the original regression problem reduces to many separate small-size regression problems. In detail, the SAS property implies that $\mathcal{U}_p(t_p^*)$ splits into many connected subgraphs, each is small in size, and different ones are disconnected. Given two disjoint connected subgraphs $\mathcal{I}_0$ and $\mathcal{J}_0$ where $\mathcal{I}_0\lhd\mathcal{U}_p(t)$ and $\mathcal{J}_0\lhd\mathcal{U}_p(t)$,

$$\Omega^*(i,j)=0\qquad \forall i\in\mathcal{I}_0, j\in\mathcal{J}_0. \tag{2.12}$$

Recall that the regression model (1.1) is closely related to the model $X'Y=X'X\beta+X'z$. Fixing a connected subgraph $\mathcal{I}_0\lhd\mathcal{U}_p(t_p^*)$, we restrict our attention to $\mathcal{I}_0$ by considering $(X'Y)^{\mathcal{I}_0}=(X'X\beta)^{\mathcal{I}_0}+(X'z)^{\mathcal{I}_0}$. See Definition 1.1 for notation. Since $X_i\overset{\text{i.i.d.}}{\sim}N(0,\frac{1}{n}\Omega)$ and $\mathcal{I}_0$ has a small size, we expect to see $(X'X\beta)^{\mathcal{I}_0}\approx(\Omega\beta)^{\mathcal{I}_0}$ and $(X'z)^{\mathcal{I}_0}\approx N(0,\Omega^{\mathcal{I}_0,\mathcal{I}_0})$. Therefore, $(X'Y)^{\mathcal{I}_0}\approx N((\Omega\beta)^{\mathcal{I}_0},\Omega^{\mathcal{I}_0,\mathcal{I}_0})$. A key observation is

$$(\Omega\beta)^{\mathcal{I}_0}\approx\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}. \tag{2.13}$$

In fact, letting $\mathcal{I}_0^c=\{j:1\le j\le p, j\notin\mathcal{I}_0\}$, it is seen that

$$(\Omega\beta)^{\mathcal{I}_0}-\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}=(\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}+(\Omega-\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}=\mathrm{I}+\mathrm{II}. \tag{2.14}$$

First, by Lemma 2.2, $|\mathrm{II}|\le C\|\Omega-\Omega^*\|_\infty\|\beta\|_\infty=o(\sqrt{\log(p)})$ coordinate-wise, hence II is negligible. Second, by the sure screening property, signals that are

falsely screened out in the U-step are fewer than $L_p p^{1-(\vartheta+r)^2/(4r)}$, and therefore have a negligible effect. To bring out the intuition, we assume $\mathcal{U}_p(t_p^*)$ contains all signals for a moment (see [18], Lemma A.4, for formal treatment). This, with (2.12), implies that $\mathrm{I} = 0$, and (2.13) follows.

As a result, the original regression problem reduces to many small-size regression problems of the form

$$(X'Y)^{\mathcal{I}_0} \approx N(\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}, \Omega^{\mathcal{I}_0,\mathcal{I}_0}) \tag{2.15}$$

that can be fitted separately. Note that $\Omega^{\mathcal{I}_0,\mathcal{I}_0}$ can be accurately estimated by $(X'X)^{\mathcal{I}_0,\mathcal{I}_0}$, due to the small size of $\mathcal{I}_0$. We are now ready for the P-step.

2.4. *P-step.* The goal of the P-step is that, for each fixed connected subgraph $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$, we fit model (2.15) with an error rate $\le L_p p^{-(\vartheta+r)^2/(4r)}$. This turns out to be rather delicate, and many methods (including the lasso and the subset selection) do not achieve the desired rate of convergence.

For this reason, we proposed a penalized-MLE approach. The idea can be explained as follows. Given that $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$ as a priori, the chance that $\mathcal{I}_0$ contains k signals is $\sim \varepsilon_p^k$. This motivates us to fit model (2.15) by maximizing the likelihood function $\varepsilon_p^k \cdot \exp[-\frac{1}{2}[(X'Y)^{\mathcal{I}_0} - A\mu]'A^{-1}[(X'Y)^{\mathcal{I}_0} - A\mu]]$, subject to $\|\mu\|_0 = k$. Recalling $A = (X'X)^{\mathcal{I}_0,\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}$, this is proportional to the density of $(X'Y)^{\mathcal{I}_0}$ in (2.15), hence the name of penalized MLE. Recalling $\varepsilon_p = p^{-\vartheta}$ and $\lambda_p^{\mathrm{ups}} = \sqrt{2\vartheta \log p}$, it is equivalent to minimizing

$$[(X'Y)^{\mathcal{I}_0} - A\mu]'A^{-1}[(X'Y)^{\mathcal{I}_0} - A\mu] + (\lambda_p^{\mathrm{ups}})^2 \cdot \|\mu\|_0. \tag{2.16}$$

Unfortunately, (2.16) does not achieve the desired rate of convergence as expected. The reason is that we have not taken full advantage of the information provided: given that all coordinates in $\mathcal{I}_0$ survive the screening, each signal in $\mathcal{I}_0$ should be relatively strong. Motivated by this, for some tuning parameter $u^{\mathrm{ups}} > 0$, we force all nonzero coordinates of μ to equal u^{ups}. This is the UPS procedure we introduced in Section 1. In Theorem 2.1 below, we show that this procedure obtains the desired rate of convergence provided that u^{ups} is properly set.

One may think that forcing all nonzero coordinates of μ to be equal is too restrictive, since the nonzero coordinates of $\beta^{\mathcal{I}_0}$ are unequal. Nevertheless, the UPS achieves the desired error rate. The reason is that, knowing the exact values of the nonzero coordinates is not crucial, as the main goal is to separate nonzero coordinates of $\beta^{\mathcal{I}_0}$ from the zero ones.

Similarly, since knowing the signal distribution π_p may be very helpful, one may choose to estimate π_p using the data first and then combine the estimated distribution with the P-step. However, this has two drawbacks. First, model (2.15) is very small in size, and can be easily over fit if we introduce too many degrees of freedom. Second, estimating π_p usually involves deconvolution, which generally has relatively slow rate of convergence (e.g., [26]); a noisy estimate of π_p may hurt rather than help in fitting model (2.15).

2.5. *Upper bound.* We are now ready for the upper bound. To recap, the proposed procedure is as follows:

- With fixed tuning parameters $(t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}})$, obtain $\mathcal{U}_p(t) = \{j : 1 \le j \le p, (x_j, Y) \ge t\}$.
- Obtain Ω^* as in (2.11), and form a graph (V_0, Ω_0) with $V_0 = \{1, 2, \ldots, p\}$ and $\Omega_0 = \Omega^*$.
- Split $\mathcal{U}_p(t)$ into connected subgraphs where different ones are disconnected. For each connected subgraph $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\}$, obtain the minimizer of (2.16), where each coordinate of μ is either 0 or u^{ups}. Denote the estimate by $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}}, p)$.
- For any $1 \le j \le p$, if $j \notin \mathcal{U}_p(t)$, set $\hat{\beta}_j = 0$. Otherwise, there is a unique $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \lhd \mathcal{U}_p(t)$, where $i_1 < i_2 < \cdots < i_K$, such that j is the kth coordinate of $\mathcal{I}_0$. Set $\hat{\beta}_j = (\hat{\mu}(\mathcal{I}_0))_k$.

Denote the resulting estimator by $\hat{\beta}(Y, X; t, \lambda^{\mathrm{ups}}, u^{\mathrm{ups}})$. We have the following theorem.

THEOREM 2.1. *Consider model (2.1)–(2.2) where (2.3)–(2.7) hold, and fix $0 < q \le (\vartheta + r)^2/(4r)$. For sufficiently large p, if $\Omega^{(p)} \in \mathcal{M}_p^+(\omega_0, \gamma, A)$, and we set the tuning parameters of the UPS at*

$$t = t_p^* = \sqrt{2q\log(p)}, \qquad \lambda^{\mathrm{ups}} = \lambda_p^{\mathrm{ups}} = \sqrt{2\vartheta \log p}, \qquad u^{\mathrm{ups}} = u_p^{\mathrm{ups}} = \tau_p,$$

then as $p \to \infty$, $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ups}}(Y, X; t_p^*, \lambda_p^{\mathrm{ups}}, u_p^{\mathrm{ups}}), \vartheta, r, \Omega^{(p)}) \le L_p \cdot s_p \cdot p^{-(r-\vartheta)^2/(4r)}$. The claim remains valid if $r/\vartheta \le 3 + 2\sqrt{2}$ and $\Omega^{(p)} \in \mathcal{M}_p^*(\omega_0, \gamma, A)$ for sufficiently large p.

Except for the L_p term, the upper bound matches the lower bound in Theorem 1.1. Therefore, both bounds are tight and the UPS is rate optimal.

2.6. *Tuning parameters of the UPS.* The UPS uses three tuning parameters $(t_p^*, \lambda_p^{\mathrm{ups}}, u_p^{\mathrm{ups}})$. In this section, we show that under certain conditions, the parameters $(\lambda_p^{\mathrm{ups}}, u_p^{\mathrm{ups}})$ can be estimated from the data.

In detail, recall that $\tilde{Y} = X'Y$. For $t > 0$, introduce $\bar{F}_p(t) = \frac{1}{p}\sum_{j=1}^p 1\{\tilde{Y}_j > t\}$ and $\mu_p(t) = \frac{1}{p}\sum_{j=1}^p \tilde{Y}_j \cdot 1\{\tilde{Y}_j > t\}$. Denote the largest off-diagonal coordinate of Ω by $\delta_0 = \delta_0(\Omega) = \max_{\{1 \le i,j \le p, i \ne j\}} |\Omega(i,j)|$. Recalling that the support of π_p is contained in $[\tau_p, (1+\eta)\tau_p]$, we suppose

$$2\delta_0(1+\eta) - 1 \le \vartheta/r \qquad \text{so that } \delta_0^2(1+\eta)^2 r < \frac{(\vartheta + r)^2}{4r}. \tag{2.17}$$

Let $\mu_p^*(\pi_p)$ be the mean of π_p. The following is proved in [18].

LEMMA 2.4. *Fix q such that* $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \le (\vartheta + r)^2/(4r)$, *and let* $t_p^* = \sqrt{2q\log p}$. *Suppose the conditions in Theorem 2.1 hold. As* $p \to \infty$, *with probability of* $1 - o(1/p)$,

$$|[\bar{F}_p(t_p^*)/\varepsilon_p] - 1| = o(1) \quad \text{and} \quad |[\mu_p(t_p^*)/(\varepsilon_p \mu_p^*(\pi_p))] - 1| = o(1). \tag{2.18}$$

Motivated by Lemma 2.18, we propose to estimate $(\lambda^{\mathrm{ups}}, u^{\mathrm{ups}})$ by

$$\begin{aligned} \hat{\lambda}_p^{\mathrm{ups}} &= \hat{\lambda}_p^{\mathrm{ups}}(q) = \sqrt{-2\log(\bar{F}_p(t_p^*))}, \\ \hat{u}_p^{\mathrm{ups}} &= \hat{u}_p^{\mathrm{ups}}(q) = \mu_p(t_p^*)/\bar{F}_p(t_p^*). \end{aligned} \tag{2.19}$$

THEOREM 2.2. *Fix q such that* $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \le (\vartheta + r)^2/(4r)$, *and let* $t_p^* = \sqrt{2q\log p}$. *Suppose the conditions of Theorem 2.1 hold. As* $p \to \infty$, *if additionally* $\mu_p^*(\pi_p) \le (1 + o(1))\tau_p$, *then* $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ups}}) \le L_p \cdot s_p \cdot p^{-(r-\vartheta)^2/(4r)}$.

As a result, t_p^* is the only tuning parameter needed by the UPS. By Theorem 2.2, the performance of the UPS is relatively insensitive to the choice of t_p^*, as long as it falls in a certain range. Numerical studies in Section 5 confirm this for finite p. The numerical study also suggests that the lasso is comparably more sensitive to its tuning parameter λ^{lasso}.

2.7. *Discussions.* While the conditions in Theorems 2.1 are 2.2 are relatively strong, the key idea of the paper applies to much broader settings. The success of UPS attributes to the interaction of the signal sparsity and graph sparsity, which can be found in many applications [e.g., compressive sensing, genome-wide association study (GWAS)].

In the forthcoming papers [11, 19, 20], we revisit the key idea of this paper, and extend our results to more general settings. However, the current paper is different from [11, 19, 20] in important ways. First, the focus of [11] is on ill-posed regression models and change-point problems, and the focus of [20] is on Ising model and network data. Second, the current paper uses the so-called "phase diagram" as a new criterion for optimality (e.g., [10]), and Jin and Zhang [19] use the more traditional "asymptotic minimaxity" as the criterion for optimality. Due to the complexity of the problem, one type of optimality usually does not imply the other. The current paper and [19] have very different targets, objectives and underlying mathematical techniques, and the results in either one cannot be deduced from the other.

The current paper is new in at least two aspects. First, given that marginal regression is a widely used method but is not well justified, this paper shows that marginal regression can actually work, provided that an additional cleaning stage is performed. Second, it shows that L^0-penalization method—the target of many relaxation methods—is nonoptimal, even in very simple settings, and even when the tuning parameter is ideally set.

3. A refinement for moderately large p. We introduce a refinement for the UPS when p is moderately large. We begin by investigating the relationship between the regression model and Stein's normal means model.

Recall that model (1.1) is closely related to the following model:

$$X'Y = X'X\beta + X'z, \qquad z \sim N(0, I_n), \tag{3.1}$$

which is approximately equivalent to Stein's normal means model as follows:

$$X'Y \approx \Omega\beta + N(0,\Omega) \quad \Longleftrightarrow \quad \Omega^{-1}X'Y \approx N(\beta, \Omega^{-1}). \tag{3.2}$$

In the literature, Stein's normal means model has been extensively studied, but the focus has been on the case where Ω is diagonal (e.g., [26]). When Ω is not diagonal, Stein's normal means model is intrinsically a regression problem. To see how close models (3.1) and (3.2) are, write

$$X'Y = \left[\Omega\beta + \frac{\sqrt{n}}{\|z\|}X'z\right] + \left[(X'X-\Omega)\beta + \left(\frac{\|z\|}{\sqrt{n}}-1\right)\frac{\sqrt{n}}{\|z\|}X'z\right] = \mathrm{I} + \mathrm{II}. \tag{3.3}$$

First, note that $\mathrm{I} \sim N(\Omega\beta, \Omega)$. For II, we have the following lemma.

LEMMA 3.1. *Consider model (2.1)–(2.2) where (2.2)–(2.4) hold. As $p \to \infty$, there is a constant $C > 0$ such that except for a probability of $o(1/p)$,*

$$\left|\frac{\|z\|}{\sqrt{n}} - 1\right| \le C(\sqrt{\log p})p^{-\theta/2},$$

$$\|(X'X-\Omega)\beta\|_\infty \le C\|\Omega\|(\sqrt{2\log p})p^{-(\theta-(1-\vartheta))/2}.$$

It follows that $|\mathrm{II}| \le C\sqrt{2\log(p)}\cdot p^{-[\theta-(1-\vartheta)]/2}$ coordinate-wise. Therefore, *asymptotically*, models (3.1) and (3.2) have negligible difference. However, when p is moderately large, the difference between models (3.1) and (3.2) may be nonnegligible. In Table 1, we tabulate the values of $\sqrt{2\log(p)}\cdot p^{-[\theta-(1-\vartheta)]/2}$, which are relatively large for moderately large p.

This says that, for moderately large p, the random design model is much noisier than Stein's normal means model. As a result, in the U-step, we tend to falsely keep more noise terms in the former than in the latter; some of these noise terms are large in magnitude, and it is hard to clean all of them

TABLE 1
The values of $\sqrt{2\log(p)}p^{-[\theta-(1-\vartheta)]/2}$ for different p and (θ,ϑ)

p	400	5×400	$5^2\times 400$	$5^3\times 400$	$5^4\times 400$	$5^5\times 400$
$(\theta,\vartheta)=(0.91,0.65)$	0.65	0.46	0.33	0.22	0.15	0.10
$(\theta,\vartheta)=(0.91,0.5)$	1.01	0.82	0.65	0.51	0.39	0.30

in the P-step. To see how the problem can be fixed, we write

$$X'X\beta = (X'X - \Omega^*)\beta + \Omega^*\beta. \tag{3.4}$$

On one hand, the term $(X'X - \Omega^*)\beta$ causes the random design model to be much noisier than Stein's normal means model. On the other hand, this term can be easily removed from the model if we have a reasonably good estimate of β. This motivates a refinement as follows.

For any $p \times 1$ vector y, let $S^2(y) = \frac{1}{p-1}\sum_{j=1}^p (y_j - \bar{y})^2$ where $\bar{y} = \frac{1}{p}\sum_{j=1}^p y_j$. We propose the following procedure: (1) Run the UPS and obtain an estimate of β, say, $\hat{\beta}$. Let $W^{(0)} = X'Y$ and $\hat{\beta}^{(0)} = \hat{\beta}$. (2) For $j = 1, 2, 3$, respectively, let $W^{(j)} = X'Y - (X'X - \Omega^*)\hat{\beta}^{(j-1)}$. If $S(W^{(j)})/S(W^{(j-1)}) \leq 1.05$, run the UPS with $X'Y$ replaced by $W^{(j)}$ and other parts unchanged, and let $\hat{\beta}^{(j)}$ be the new estimate. Stop otherwise.

Numerical studies in Section 5 suggest that the refinement is beneficial for moderately large p. When p is sufficiently large [e.g., $\sqrt{2\log(p)} \cdot p^{-[\theta-(1-\vartheta)]/2} \leq 0.4$], the original UPS is usually good enough. In this case, refinements are not necessary, but may still offer improvements.

4. Understanding the lasso and the subset selection. In this section, we show that there is a region in the phase space where the lasso is rate nonoptimal (similarly for subset selection). We use Stein's normal means model instead of the random design model (as the goal is to understand the nonoptimality of these methods, focusing on a simpler model enjoys mathematical convenience, yet is also sufficient; see Section 1.7).

To recap, the model we consider in this section is $\tilde{Y} \sim N(\Omega\beta, \Omega)$, where $\tilde{Y}$ is the counterpart of $X'Y$ in the random design model. Fix $a \in (-1/2, 1/2)$. As in Section 1.7, we let Ω be the tridiagonal matrix as in (1.12), and π_p be the point mass at $\tau_p = \sqrt{2r\log p}$. In other words,

$$\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p\nu_{\tau_p}, \qquad \varepsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}. \tag{4.1}$$

Throughout this section, we assume $r > \vartheta$ so that successful variable selection is possible. Somewhat surprisingly, even in this simple case and even when (ε_p, τ_p) are known, there is a region in the phase space where neither the lasso nor the subset selection is optimal. To shed light, we first take a heuristic approach below. Formal statements are given later.

4.1. *Understanding the lasso.* The vector $\tilde{Y}$ consists of three main components: true signals, fake signals and pure noise (see Definition 1.3). According to (4.1), true signals may appear as singletons, pairs, triplets, etc., but singletons are the most common and therefore have the major effect. For each signal singleton, since Ω is tridiagonal, we have two fake signals, one to the left and one to the right. Given a site j, $1 \leq j \leq p$, the lasso may make three types of errors:

- *Type* I. $\tilde{Y}_j$ is a pure noise, but the lasso mistakes it as a signal.
- *Type* II. $\tilde{Y}_j$ is a signal singleton, but the lasso mistakes it as a noise.
- *Type* III. $\tilde{Y}_j$ is a fake signal next to a signal singleton, but the lasso mistakes it as a signal.

There are other types of errors, but these are the major ones.

To minimize the sum of these errors, the lasso needs to choose the tuning parameter λ^{lasso} carefully. To shed light, we first consider the uncorrelated case where Ω is the identity matrix. In this case, we do not have fake signals and it is understood that the lasso is equivalent to the soft-thresholding procedure [26], where the expected sum of types I and II errors is

$$p[(1-\varepsilon_p)\bar{\Phi}(\lambda^{\mathrm{lasso}}) + \varepsilon_p \Phi(\lambda^{\mathrm{lasso}} - \tau_p)]. \tag{4.2}$$

Here, $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. In (4.2), fixing $0 < q < 1$ and taking $\lambda^{\mathrm{lasso}} = \lambda_p^{\mathrm{lasso}} = \sqrt{2q\log(p)}$, the expected sum of errors is

$$\sim \begin{cases} L_p[p^{1-q} + p^{1-(\vartheta+(\sqrt{q}-\sqrt{r})^2)}], & \text{if } 0 < q < r, \\ p^{1-q} + p^{1-\vartheta}, & \text{if } q > r. \end{cases}$$

The right-hand side is minimized at $q = (\vartheta + r)^2/(4r)$ at which $\lambda_p^{\mathrm{lasso}} = \frac{\vartheta+r}{2r}\tau_p$, and the sum of errors is $L_p p^{1-(\vartheta+r)^2/(4r)}$, which is the optimal rate of convergence. For a smaller q, the lasso keeps too many noise terms. For a larger q, the lasso kills too many signals.

Return to the correlated case. The vector $\tilde{Y}$ is at least as noisy as that in the uncorrelated case. As a result, to control the type I errors, we should choose $\lambda_p^{\mathrm{lasso}}$ to be at least $\frac{\vartheta+r}{2r}\tau_p$. This is confirmed in Lemma 4.2 below.

In light of this, we fix $q \geq (\vartheta + r)^2/(4r)$ and let $\lambda_p^{\mathrm{lasso}} = \sqrt{2q\log(p)}$ from now on. We observe that except for a negligible probability, the support of $\hat{\beta}^{\mathrm{lasso}}$, denoted by $\hat{S}_p^{\mathrm{lasso}}$, splits into many small clusters (i.e., block of adjacent indices). There is an integer K not depending on p that has the following effects: (a) If $\tilde{Y}_j$ is a pure noise, and there is no signal within a distance of K from it, then either $\hat{\beta}_j^{\mathrm{lasso}} = 0$, or $\hat{\beta}_j^{\mathrm{lasso}} \neq 0$ but $\hat{\beta}_{j\pm1}^{\mathrm{lasso}} = 0$, and (b) If $\tilde{Y}_j$ is a signal singleton, and there is no other signal within a distance of K from it, then either $\hat{\beta}_j^{\mathrm{lasso}} = 0$, or $\hat{\beta}_j^{\mathrm{lasso}} \neq 0$ but $\hat{\beta}_{j\pm2} = 0$ and at least one of $\{\hat{\beta}_{j+1}^{\mathrm{lasso}}, \hat{\beta}_{j-1}^{\mathrm{lasso}}\}$ is 0. These heuristics are justified in [17] (we use such heuristics to provide insight, but not for proving results below).

At the same time, let $\mathcal{I}_0 = \{j-k+1, \ldots, j\} \subset \hat{S}_p^{\mathrm{lasso}}$ be a cluster, so that $\hat{\beta}_{j-k}^{\mathrm{lasso}} = \hat{\beta}_{j+1}^{\mathrm{lasso}} = 0$. Since Ω is tridiagonal, $(\hat{\beta}^{\mathrm{lasso}})^{\mathcal{I}_0}$, the restriction of $\hat{\beta}^{\mathrm{lasso}}$ to $\mathcal{I}_0$, is the solution of the following small-size minimization problem:

$$\tfrac{1}{2}\mu'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})\mu - \mu'\tilde{Y}^{\mathcal{I}_0} + \lambda^{\mathrm{lasso}}\|\mu\|_1 \qquad \text{where } \mu \text{ is a } k \times 1 \text{ vector.} \tag{4.3}$$

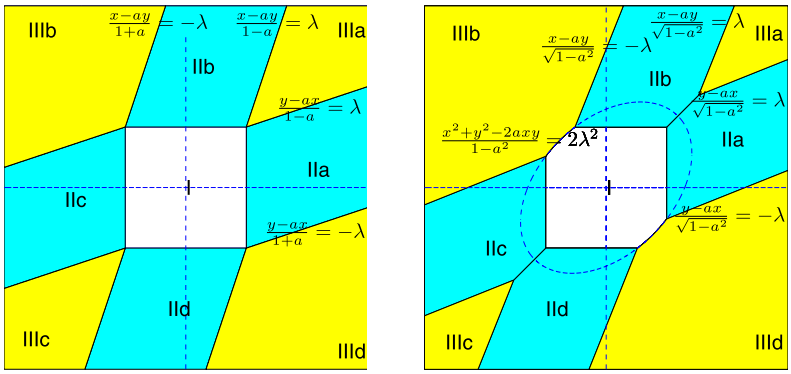

FIG. 3. *Partition of regions as in Lemma 4.1 (left) and in Lemma 4.3 (right).*

See Definition 1.1. Two special cases are noteworthy. First, $\mathcal{I}_0 = \{j\}$, and the solution of (4.3) is given by $\hat{\beta}_j^{\text{lasso}} = \text{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda^{\text{lasso}})^+$, which is the soft-thresholding [26]. Second, $\mathcal{I}_0 = \{j-1, j\}$. We call the solution of (4.3) in this case the *bivariate lasso*. We have the following lemma, where all regions I-IIId are illustrated in Figure 3 (x-axis is $\tilde{Y}_{j-1}$, y-axis is $\tilde{Y}_j$).

LEMMA 4.1. *Denote* $\lambda = \lambda^{\text{lasso}}$. *The solution of the bivariate lasso* $(\hat{\beta}_{j-1}^{\text{lasso}}, \hat{\beta}_j^{\text{lasso}})$ *is given by* $(\hat{\beta}_{j-1}^{\text{lasso}}, \hat{\beta}_j^{\text{lasso}}) = (\text{sgn}(\tilde{Y}_{j-1})(|\tilde{Y}_{j-1}| - \lambda)^+, \text{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda)^+)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* I, IIa-IId *and* $(\hat{\beta}_{j-1}^{\text{lasso}}, \hat{\beta}_j^{\text{lasso}}) = \frac{1}{1-a^2}(Z_{j-1} - aZ_j, Z_j - aZ_{j-1})$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIIa-IIId. *Here,* $Z_{j-1} = \tilde{Y}_{j-1} - \lambda$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIIa, IIId *and* $Z_{j-1} = \tilde{Y}_{j-1} + \lambda$ *otherwise;* $Z_j = \tilde{Y}_j - \lambda$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIIa, IIIb *and* $Z_j = \tilde{Y}_j + \lambda$ *otherwise.*

In the white region of Figure 3, both $\hat{\beta}_{j-1}^{\text{lasso}}$ and $\hat{\beta}_j^{\text{lasso}}$ are 0. In the blue regions, exactly one of them is 0. In the yellow regions, both are nonzero. Lemma 4.1 is proved in [18].

As a result, the following hold, except for a negligible probability:

- *Type* I. There are $O(p)$ indices j where $\tilde{Y}_j$ is a pure noise, and no signal appears within a distance of K from it. For each of such j, the lasso acts on $\tilde{Y}_j$ as (univariate) soft-thresholding, and $\hat{\beta}_j^{\text{lasso}} \neq 0$ if and only if $|\tilde{Y}_j| \geq \lambda_p^{\text{lasso}}$.
- *Types* II–III. There are $O(p\varepsilon_p)$ indices where $\tilde{Y}_j$ is a signal singleton, and no other signal appears within a distance of K from it. The lasso either acts on $\tilde{Y}_j$ as soft-thresholding, or acts on both $\tilde{Y}_j$ and one of its neighbors as the bivariate lasso. As a result, $\hat{\beta}_j^{\text{lasso}} = 0$ if and only if $|\tilde{Y}_j| \leq \lambda_p^{\text{lasso}}$ (type II), and both $\hat{\beta}_j^{\text{lasso}}$ and $\hat{\beta}_{j-1}^{\text{lasso}}$ are nonzero if and only if $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ falls in regions IIIa-IIId, with IIIa and IIIb being the most likely (type III).

Noting that $\tilde{Y}_j \sim N(0,1)$ if it is a pure noise and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton, the sum of types I and II errors is $L_p p[P(N(0,1) \geq \lambda_p^{\mathrm{lasso}}) + \varepsilon_p P(N(\tau_p,1) < \lambda_p^{\mathrm{lasso}})] = L_p p[\bar{\Phi}(\lambda_p^{\mathrm{lasso}}) + \varepsilon_p \Phi(\lambda_p^{\mathrm{lasso}} - \tau_p)]$. Also, when $\tilde{Y}_j$ is a signal singleton, $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ is distributed as a bivariate normal with means $a\tau_p$ and τ_p, variances 1, and correlation a. Denote such a bivariate normal distribution by W for short. The type III error is $L_p p \cdot P(\beta_{j-1} = 0, \beta_j = \tau_p, \ (\tilde{Y}_{j-1}, \tilde{Y}_j)' \in \text{regions IIIa or IIIb}) \sim L_p p \varepsilon_p \cdot P(W \in \text{regions IIIa or IIIb})$. Therefore, the sum of three types of errors is

$$L_p p \cdot [\bar{\Phi}(\lambda_p^{\mathrm{lasso}}) + \varepsilon_p \Phi(\lambda_p^{\mathrm{lasso}} - \tau_p) + \varepsilon_p P(W \in \text{regions IIIa or IIIb})], \tag{4.4}$$

which can be conveniently evaluated. Note that the sum of types I and II errors in the correlated case is the same as that in the uncorrelated case, which is minimized at $\lambda_p^{\mathrm{lasso}} = (\vartheta + r)/(2r)\tau_p$. Therefore, whether the lasso is optimal or not depends on whether the type III error is smaller than the optimal rate of convergence or not. Unfortunately, in certain regions of the phase space, the type III error can be significantly larger than the optimal rate. In other words, provided that the tuning parameters are properly set, the lasso is able to separate the signal singletons from the pure noise. However, it may not be efficient in filtering out the fake signals, which is the culprit for its nonoptimality.

For short, write $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{lasso}}(\lambda_p^{\mathrm{lasso}})) = \mathrm{Hamm}(\hat{\beta}^{\mathrm{lasso}}(\lambda_p^{\mathrm{lasso}}); \varepsilon_p, \tau_p, a)$. The following is proved in [18], confirming the above heuristics.

LEMMA 4.2. *Fix $\vartheta \in (0,1)$, $r > \vartheta$, $q > 0$ and $a \in (-1/2, 1/2)$. Set the lasso tuning parameter as $\lambda_p^{\mathrm{lasso}} = \sqrt{2q \log p}$. As $p \to \infty$,*

$$\frac{\mathrm{Hamm}(\hat{\beta}^{\mathrm{lasso}}(\lambda_p^{\mathrm{lasso}}))}{s_p} \geq \begin{cases} L_p p^{-\min\{((1-|a|)/(1+|a|))q, q-\vartheta\}}, & \text{if } 0 < q < \dfrac{(\vartheta+r)^2}{4r}, \\ L_p p^{-\min\{((1-|a|)/(1+|a|))q, (\sqrt{r}-\sqrt{q})^2\}}, & \text{if } \dfrac{(\vartheta+r)^2}{4r} < q < r, \\ (1+o(1)), & \text{if } q > r. \end{cases}$$

The exponent on the right-hand side is minimized at $q = (\vartheta + r)^2/(4r)$ when $r < [(1+\sqrt{1-a^2})/|a|]\vartheta$ and $q = (1+|a|)(1-\sqrt{1-a^2})r/(2a^2)$ when $r > [(1+\sqrt{1-a^2})/|a|]\vartheta$, where we note that $r < [(1+\sqrt{1-a^2})/|a|]\vartheta$ and $r > [(1+\sqrt{1-a^2})/|a|]\vartheta$ correspond to the optimal and nonoptimal regions of the lasso, respectively. This shows that in the optimal region of the lasso, $\lambda_p^{\mathrm{lasso}} = (\vartheta + r)/(2r)\tau_p$ remains the optimal tuning parameter, at which the sum of types I and II errors is minimized, and the type III error has a negligible

effect. In the nonoptimal region of the lasso, at $\lambda_p^{\mathrm{lasso}} = (\vartheta + r)/(2r)\tau_p$, the type III error is larger than the sum of types I and II errors, so the lasso needs to raise the tuning parameter slightly to minimize the sum of all three types of errors (but the resultant Hamming error is still larger than that of the optimal procedure). Combining this with Lemma 4.2 gives the following theorem, the proof of which is omitted.

THEOREM 4.1. *Set* $\lambda_p^{\mathrm{lasso}} = \sqrt{2q\log p}$. *For all choices of* $q > 0$, *the error rate of the lasso satisfies* $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{lasso}}(\lambda_p^{\mathrm{lasso}})) \geq L_p \cdot s_p \cdot p^{-(\vartheta - r)^2/(4r)}$ *when* $r/\vartheta < (1 + \sqrt{1-a^2})/|a|$ *and*

$$\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{lasso}}(\lambda_p^{\mathrm{lasso}})) \geq L_p \cdot s_p \cdot p^{\vartheta - ((1-|a|)(1-\sqrt{1-a^2})/(2a^2))r},$$

when $r/\vartheta > (1 + \sqrt{1-a^2})/|a|$.

In [17], we show that when $r/\vartheta \leq 3 + 2\sqrt{2}$, the lower bound in Theorem 4.1 is tight. The proofs are relatively long, so we leave the details to [17].

4.2. *Understanding subset selection.* The discussion is similar, so we keep it brief. Fix $1 \leq j \leq p$. The major errors that subset selection makes are the following (type III is defined differently from that in the preceding section):

- *Type* I. $\tilde{Y}_j$ is a pure noise, but subset selection takes it as a signal.
- *Type* II. $\tilde{Y}_j$ is a signal singleton, but subset selection takes it as a noise.
- *Type* III. $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is a signal pair, but subset selection mistakes one of them as a noise.

Suppose that $\tilde{Y}_j$ is either a pure noise or a signal singleton, and for an appropriately large K, no other signal appears within a distance of K from it. In this case, except for a negligible probability, $\hat{\beta}_{j\pm 1}^{\mathrm{lasso}} = 0$, and the subset selection acts on site j as hard thresholding [26], $\hat{\beta}_j^{\mathrm{ss}} = \tilde{Y}_j \cdot 1\{|\tilde{Y}_j| \geq \lambda^{\mathrm{ss}}\}$. Recall that $\tilde{Y}_j \sim N(0,1)$ if it is a pure noise, and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton. Take $\lambda^{\mathrm{ss}} = \lambda_p^{\mathrm{ss}} = \sqrt{2q\log p}$ as before. Similarly, the expected sum of types I and II errors is

$$\begin{aligned} &L_p p[\bar{\Phi}(\lambda_p^{\mathrm{ss}}) + p^{-\vartheta}\Phi(\lambda_p^{\mathrm{ss}} - \tau_p)] \\ &\quad = \begin{cases} L_p(p^{1-q} + p^{1-\vartheta-(\sqrt{q}-\sqrt{r})^2}), & \text{if } 0 < q < r, \\ L_p(p^{1-q} + p^{1-\vartheta}), & \text{if } q > r. \end{cases} \end{aligned} \tag{4.5}$$

On the right-hand side, the exponent is minimized at $q = (\vartheta + r)^2/4r$, at which the rate is $L_p p^{1-(\vartheta+r)^2/(4r)}$, which is the optimal rate of convergence.

Next, consider the type III error. Suppose $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is a signal pair and no other signal appears within a distance of K for a properly large K.

Similarly, since Ω is tridiagonal, $(\hat{\beta}^{\mathrm{ss}}_{j-1}, \hat{\beta}^{\mathrm{ss}}_j)'$ is the minimizer of the functional $\frac{1}{2}\beta_{j-1}^2 + \frac{1}{2}\beta_j^2 + a\beta_{j-1}\beta_j - (\tilde{Y}_{j-1}\beta_{j-1} + \tilde{Y}_j\beta_j) + \frac{(\lambda_p^{\mathrm{ss}})^2}{2}(I\{\beta_{j-1} \neq 0\} + I\{\beta_j \neq 0\})$. We call the resultant procedure *bivariate subset selection*. The following lemma is proved in [18], with the regions illustrated in Figure 3.

LEMMA 4.3. *The solution of the bivariate subset selection is given by* $(\hat{\beta}^{\mathrm{ss}}_{j-1}, \hat{\beta}^{\mathrm{ss}}_j) = (0,0)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in region* I, $(\hat{\beta}^{\mathrm{ss}}_{j-1}, \hat{\beta}^{\mathrm{ss}}_j) = (\tilde{Y}_{j-1}, 0)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIa, IIc, $(\hat{\beta}^{\mathrm{ss}}_{j-1}, \hat{\beta}^{\mathrm{ss}}_j) = (0, \tilde{Y}_j)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIb, *i.i.d. and* $(\hat{\beta}^{\mathrm{ss}}_{j-1}, \hat{\beta}^{\mathrm{ss}}_j) = (\frac{\tilde{Y}_{j-1} - a\tilde{Y}_j}{1-a^2}, \frac{\tilde{Y}_j - a\tilde{Y}_{j-1}}{1-a^2})$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in regions* IIIa-IIId.

When $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ falls in regions I, IIa or IIb, either $\hat{\beta}^{\mathrm{ss}}_{j-1}$ or $\hat{\beta}^{\mathrm{ss}}_j$ is 0, and the subset selection makes a type III error. Note there are $O(p\varepsilon_p^2)$ signal pairs, and that $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ is jointly distributed as a bivariate normal with means $(1+a)\tau_p$, variances 1 and correlation a. The type III error is then $L_p p^{1-(2\vartheta + \min\{[(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2, 2[(\sqrt{r(1+a)} - \sqrt{q})^+]^2\}}$. Combining with (4.5) and Mills's ratio gives the sum of all three types of errors. Formally, writing for short $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ss}}(\lambda_p^{\mathrm{ss}})) = \mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ss}}(\lambda_p^{\mathrm{ss}}); \varepsilon_p, \tau_p, a)$, we have the following lemma proved in [18].

LEMMA 4.4. *Set the tuning parameter* $\lambda_p^{\mathrm{ss}} = \sqrt{2q\log p}$. *The Hamming error for the subset selection* $\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ss}}(\lambda_p^{\mathrm{ss}}))$ *is at least*

$$\begin{cases} L_p \cdot s_p \cdot p^{-\min\{q-\vartheta, \vartheta + [(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2\}}, & \text{if } 0 < q < \dfrac{(\vartheta + r)^2}{4r}, \\ L_p \cdot s_p \cdot p^{-\min\{(\sqrt{r} - \sqrt{q})^2, \vartheta + [(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2\}}, & \text{if } \dfrac{(\vartheta + r)^2}{4r} < q < r, \\ s_p \cdot (1 + o(1)), & \text{if } q > r. \end{cases}$$

The exponents on the right-hand side are minimized at $q = (\vartheta + r)^2/(4r)$ if $r/\vartheta < [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$, and at $q = [2\vartheta + r(1-a^2)]^2/[4r(1-a^2)]$ if $r/\vartheta > [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$. As a result, we have the following theorem, the proof of which is omitted.

THEOREM 4.2. *Set the tuning parameter* $\lambda_p^{\mathrm{ss}} = \sqrt{2q\log p}$. *Then for all* $q > 0$, *the Hamming error of the subset selection satisfies*

$$\frac{\mathrm{Hamm}_p(\hat{\beta}^{\mathrm{ss}}(\lambda_p^{\mathrm{ss}}))}{s_p} \geq \begin{cases} L_p p^{-(\vartheta - r)^2/(4r)}, & \text{if } \dfrac{r}{\vartheta} < \dfrac{2 - \sqrt{1-a^2}}{\sqrt{1-a^2}(1 - \sqrt{1-a^2})}, \\ L_p p^{-[2\vartheta + r(1-a^2)]^2/(4r(1-a^2)) + \vartheta}, & \text{if } \dfrac{r}{\vartheta} > \dfrac{2 - \sqrt{1-a^2}}{\sqrt{1-a^2}(1 - \sqrt{1-a^2})}. \end{cases}$$

This gives the phase diagram in Figure 2, where (ϑ, r) satisfying $r/\vartheta < [2-\sqrt{1-a^2}]/[\sqrt{1-a^2}(1-\sqrt{1-a^2})]$ defines the optimal region, and (ϑ, r) with $r/\vartheta > [2-\sqrt{1-a^2}]/[\sqrt{1-a^2}(1-\sqrt{1-a^2})]$ defines the nonoptimal region. Similar to the lasso, the subset selection is able to separate signal singletons from the pure noise provided that the tuning parameter is properly set. But the subset selection is too harsh on signal pairs, triplets, etc., which costs its rate optimality. In [17], we further show that in certain regions of the phase space, the lower bound in Theorem 4.1 is tight.

5. Simulations. We have conducted a small-scale empirical study of the performance of the UPS. The idea is to select a few interesting combinations of $(\vartheta, \theta, \pi_p, \Omega)$ and study the behavior of the UPS for finite p. Fixing $(p, \pi_p, \Omega, \vartheta, \theta)$, let $n_p = p^\theta$ and $\varepsilon_p = p^{-\vartheta}$. We investigate both the random design model and Stein's normal means model.

In the former, the experiment contains the following steps: (1) Generate a $p \times 1$ vector β by $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p \pi_p$, and an $n_p \times 1$ vector $z \sim N(0, I_{n_p})$. (2) Generate an $n_p \times p$ matrix X the rows of which are samples from $N(0, \frac{1}{n_p}\Omega)$; let $Y = X\beta + z$. (3) Apply the UPS and the lasso. For the lasso, we use the *glmnet* package by Friedman et al. [14] (Ω is assumed unknown in both procedures). (4) Repeat 1–3 for 100 independent cycles, and calculate the average Hamming distances.

In the latter, the settings are similar, except for (i) $n_p = p$, (ii) $Y \sim N(\Omega^{1/2}\beta, I_p)$ in step 2 and (iii) Ω is assumed as known in step 3 (otherwise valid inference is impossible). We include Stein's normal means model in the study for it is the idealized version of the random design model.

EXPERIMENT 1. In this experiment, we use Stein's normal means model to investigate the boundaries of the region of exact recovery by the UPS and that by the lasso. Fixing $p = 10^4$ and Ω as the tridiagonal matrix in (1.12) with $a = 0.45$, we let ϑ range in $\{0.25, 0.5, 0.65\}$, and let $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r\log p}$, where r is chosen such that $\tau_p \in \{5, 6, \ldots, 12\}$. For both procedures, we use the ideal threshold introduced in Sections 2 and 4, respectively. That is, the tuning parameters of the UPS are set as $(t_p^*, \lambda_p^{\text{ups}}, u_p^{\text{ups}}) = (\frac{\vartheta+r}{2r}\tau_p, \sqrt{2\vartheta\log(p)}, \tau_p)$, and the tuning parameter of the lasso is set as $\lambda_p^{\text{lasso}} = \max\{\frac{\vartheta+r}{2r}, (1+\sqrt{(1-a)/(1+a)})^{-1}\}\tau_p$.

The results are reported in Table 2, where the UPS outperforms consistently over the lasso, most prominently in the case of $\vartheta = 0.25$. Also, for $\vartheta = 0.25, 0.5$, or 0.65, the Hamming errors of the UPS start to fall below 1 when τ_p exceeds $8, 7$ or 7, respectively, but that of the lasso won't fall below 1 until τ_p exceeds $12, 8$ or 7, respectively. In Section 1, we show that the UPS yields exact recovery when $\tau_p > (1+\sqrt{1-\vartheta})\sqrt{2\log p}$, where the right-hand side equals $(8.01, 7.32, 7.01)$ with the current choices of (p, ϑ). The numerical results fit well with the theoretic results.

TABLE 2
Hamming errors (Experiment 1). UPS needs weaker signals for exact recovery

	τ_p	5	6	7	8	9	10	11	12
$\vartheta = 0.25$	UPS	49	11.1	1.79	0.26	0.02	0	0	0
	lasso	186.7	99.35	58.26	38.53	25.97	18.18	12.94	10.57
$\vartheta = 0.50$	UPS	10.06	2.11	0.37	0.09	0	0	0	0
	lasso	16.36	5.11	1.47	0.51	0.28	0.33	0.26	0.09
$\vartheta = 0.65$	UPS	5.49	1.29	0.33	0.06	0	0	0	0
	lasso	7.97	2.43	0.69	0.18	0.07	0.03	0.02	0.01

EXPERIMENT 2. We use a random design model where $(p, \vartheta, \theta) = (10^4, 0.65, 0.91)$, and $\tau_p \in \{1, 2, \ldots, 7\}$. The experiment contains three parts, 2a–2c. In 2a, we take Ω to be the penta-diagonal matrix $\Omega(i,j) = 1\{i = j\} + 0.4 \cdot 1\{|i-j| = 1\} + 0.1 \cdot 1\{|i-j| = 2\}$. Also, for each τ_p, we set π_p as $\mathrm{Uniform}(\tau_p - 0.5, \tau_p + 0.5)$. In 2b, we generate Ω in a way such that it has 4 nonzero off-diagonal elements on average in each row and each column, at locations randomly chosen. Also, for each τ_p, we take π_p to be $\mathrm{Uniform}(\tau_p - 1, \tau_p + 1)$. In 2c, we use a non-Gaussian design for X. In detail, first, we generate an $n \times p$ matrix M the coordinates of which are i.i.d. samples from $\mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$. Second, we generate Ω as in 2b. Last, we let $X = (1/\sqrt{n})M\Omega^{1/2}$. Also, for each τ_p, we take π_p to be the mixture of two uniform distributions $\frac{1}{2}\mathrm{Uniform}(\tau_p - 0.5, \tau_p + 0.5) + \frac{1}{2}\mathrm{Uniform}(-\tau_p - 0.5, -\tau_p + 0.5)$. In all these experiments, the tuning parameters are set the same way as in Experiment 1. The results are reported in Table 3, suggesting that the UPS outperforms the lasso almost over the whole range of τ_p.

EXPERIMENT 3. The goal of this experiment is twofold. First, we investigate the sensitivity of the UPS and the lasso with respect to their tuning parameters. Second, we investigate the refined UPS introduced in Section 3. Fix $q > 0$. For the lasso, we take $\lambda_p^{\mathrm{lasso}} = \sqrt{2q\log(p)}$. For the UPS, set the U-step tuning parameter as $t_p^* = \sqrt{2q\log(p)}$ and let the P-step tuning parameters be estimated as in (2.19). Theorem 2.2 predicts that the UPS performs well provided that $q \in (\max\{\vartheta, \delta_0^2(1+\eta)^2 r\}, (\vartheta + r)^2/(4r))$, so both

TABLE 3
Ratios between Hamming errors and $p\varepsilon_p$ (Experiment 2a–2c). Bold: UPS. Plain: lasso

τ_p	1	2	3	4	5	6	7
2a	**1.01** 1.02	**0.96** 1.04	**0.82** 0.97	**0.51** 0.64	**0.24** 0.28	**0.09** 0.10	**0.04** 0.04
2b	**1.00** 1.00	**0.98** 1.04	**0.84** 0.96	**0.55** 0.67	**0.26** 0.32	**0.10** 0.12	**0.05** 0.05
2c	**0.94** 0.95	**0.90** 0.91	**0.89** 0.95	**0.48** 0.60	**0.18** 0.27	**0.05** 0.11	**0.01** 0.03

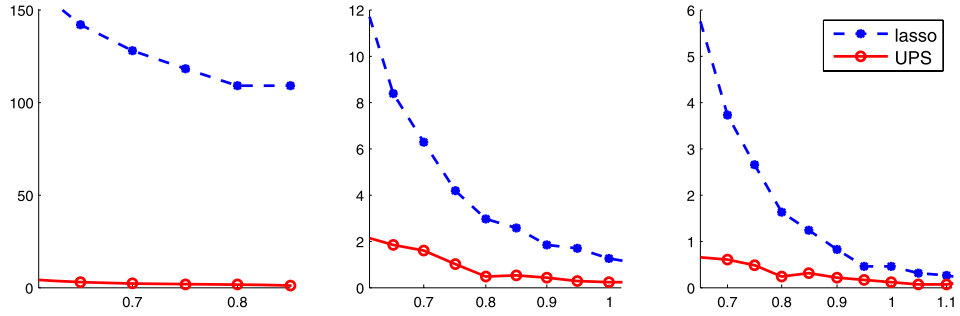

FIG. 4. *Experiment 3*a. *x-axis: q. y-axis: Hamming error. Left to right:* $\vartheta = 0.2, 0.5, 0.65$.

the lasso and the UPS are driven by one tuning parameter q. We now investigate how the choice of q affects the performances of the UPS and the lasso. The experiment contains three sub-experiments 3a–3c.

In 3a, we use Stein's normal means model where $(p, r) = (10^4, 3)$, $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r \log p}$, Ω is the penta-diagonal matrix satisfying $\Omega(i,j) = 1_{\{i=j\}} + 0.45 \cdot 1_{\{|i-j|=1\}} + 0.05 \cdot 1_{\{|i-j|=2\}}$, and $\vartheta \in \{0.2, 0.5, 0.65\}$. Note that when $\vartheta = 0.65$, $(\max\{\vartheta, \delta_0^2(1+\eta)^2 r\}, (\vartheta + r)^2/(4r)) = (0.65, 1)$ (similarly for other ϑ), so we let $q \in \{0.7, 0.8, \ldots, 1.1\}$.

In 3b, we use a random design model where (p, r, π_p, Ω, q) and the tuning parameters are the same as in 3a, but $\theta = 0.8$ and $\vartheta \in \{0.5, 0.65\}$ (the case $\vartheta = 0.2$ is relatively challenging in computation so is omitted). We compare the lasso with the refined UPS where in each iteration, we use the same tuning parameters as in 3a.

In 3c, we use the same setup as in 3b, except that we fix $q = 1$ and let τ_p range in $\{6, 6.5, \ldots, 9\}$.

The results of 3a–3c are reported in Figures 4–6, correspondingly. These results suggest that, first, the UPS consistently outperforms the lasso, and, second, the UPS is relatively less sensitive to different choices of q.

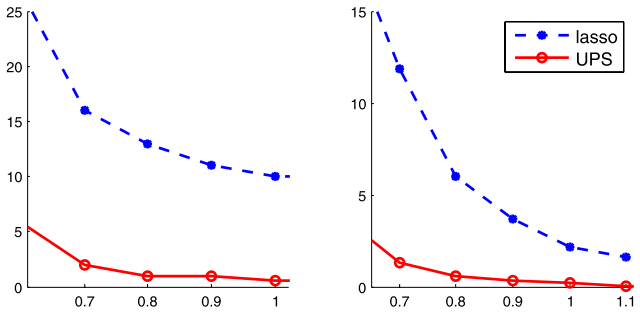

FIG. 5. *Experiment 3*b. *x-axis: q. y-axis: Hamming error. Left:* $\vartheta = 0.5$. *Right:* $\vartheta = 0.65$.

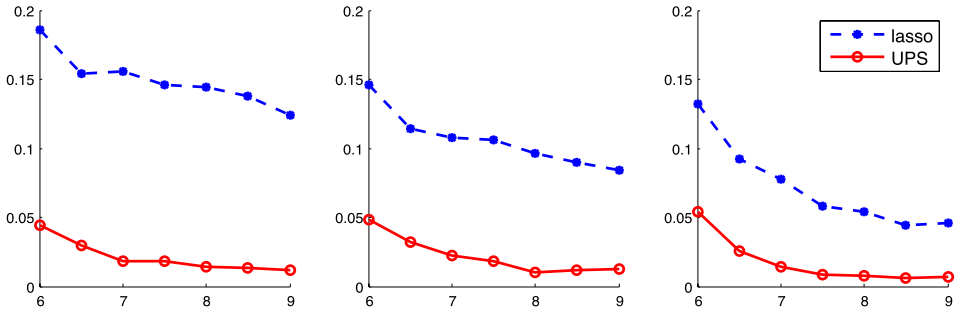

FIG. 6. *Experiment 3c. The x-axis is τ_p, and the y-axis is the ratio between the Hamming error and $p\varepsilon_p$. Left to right: $\vartheta = 0.65, 0.5, 0.2$.*

EXPERIMENT 4. In this experiment, we investigate the effect of larger p and n, respectively. The experiment includes two sub-experiments, 4a and 4b.

In 4a, we use Stein's normal means model where $(\vartheta, r) = (0.5, 3)$, Ω as in Experiment 2c, $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r\log p}$, and we let $p = 100 \times \{1, 10, 10^2, 10^3, 10^4\}$. The lasso and the UPS are implemented as in Experiment 3a, where $q = 1$. The results are reported in the left part of Table 4, where the second line displays the ratios between the Hamming errors by the lasso and that by the UPS. Theoretic results (Sections 1.7 and 4) predict that for (ϑ, r) in the nonoptimal region of the lasso, such ratios diverge as p tends to ∞. The numerical results fit well with the theory.

In 4b, we illustrate that in a random design model, if we fix p and let n increase, then the random design models get increasingly close to Stein's normal means model. In detail, we take a random design model where $(p, \vartheta, r) = (10^4, 0.5, 3)$, Ω and π_p as in Experiment 2c and $n_p = 300 \times \{1, 3, 3^2, 3^3, 3^4\}$. We also take Stein's normal means model with the same $(p, \vartheta, r, \Omega, \pi_p)$. The performance of the UPS in both models is reported in the right part of Table 4, where the last line is the ratio between the Hamming errors by the UPS for the random design model and that for the Stein's normal means model. The ratios effectively converge to 1 as n increases.

TABLE 4

Left: ratios between the Hamming errors by the UPS and that by the lasso (Experiment 4a). Right: ratios between the Hamming errors by the UPS for the random design model and that for Stein's normal means model (Experiment 4b)

p					n				
10^2	10^3	10^4	10^5	10^6	300	900	2,700	8,100	24,000
2.43	5.81	6.25	8.80	10.37	479.25	54.04	12.66	1.08	1.01

Acknowledgments. Jiashun Jin thanks Tony Cai, Emmanuel Candes, David Donoho, Stephen Fienberg, Alan Friez, Robert Nowak, Runze Li, Larry Wasserman and Cun-Hui Zhang for valuable pointers and discussion.

SUPPLEMENTARY MATERIAL

Supplementary material for "UPS delivers optimal phase diagram in high-dimensional variable selection" (DOI: 10.1214/11-AOS947SUPP; .pdf). Owing to space constraints, the technical proofs are moved to a supplementary document [18].

REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19** 716–723. MR0423716

[3] Bajwa, W. U., Haupt, J. D., Raz, G. M., Wright, S. J. and Nowak, R. D. (2007). Toeplitz-structured compressed sensing matrices. In *Proceedings of IEEE Workshop on Statistical Signal Processing* (*SSP*), *Madison*, *Wisconsin* 294–298. IEEE Computer Society, Washington, DC.

[4] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[5] Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

[6] Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

[7] Diestel, R. (2005). *Graph Theory*, 3rd ed. *Graduate Texts in Mathematics* **173**. Springer, Berlin. MR2159259

[8] Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 202–210. ACM Press, New York.

[9] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[10] Donoho, D. L. and Tanner, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102** 9446–9451 (electronic). MR2168715

[11] Fan, J., Jin, J. and Ke, Z. (2011). Optimal procedure for variable selection in the presence of strong dependence. Unpublished manuscript.

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[13] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

[14] Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22. Available at `http://cran.r-project.org/web/packages/glmnet/index.html`.

[15] Genovese, C., Jin, J. and Wasserman, L. (2011). Revisiting marginal regression. Unpublished manuscript.

[16] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357
[17] Ji, P. (2011). Selected topics in nonparametric testing and variable selection for high dimensional data. Ph.D. thesis, Dept. Statistical Science, Cornell Univ.
[18] Ji, P. and Jin, J. (2011). Supplement to "UPS delivers optimal phase diagram in high dimensional variable selection." DOI:10.1214/11-AOS947SUPP.
[19] Jin, J. and Zhang, C.-H. (2011). Adaptive optimality of UPS in high dimensional variable selection. Unpublished manuscript.
[20] Jin, J. and Zhang, Q. (2011). Optimal selection of variable when signals come from an Ising model. Unpublished manuscript.
[21] Kerkyacharian, G., Mougeot, M., Picard, D. and Tribouley, K. (2009). Learning out of leaders. In *Multiscale, Nonlinear and Adaptive Approximation* 295–324. Springer, Berlin.
[22] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[23] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[25] Wainwright, M. (2006). Sharp threshold for high-dimensional and noisy recovery of sparsity. Technical report, Dept. Statistics, Univ. California, Berkeley.
[26] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729
[27] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689
[28] Ye, F. and Zhang, C. H. (2009). Rate minimaxity of the lasso and Dantzig estimators. Technical report, Dept. Statistics and Biostatistics, Rutgers Univ.
[29] Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. Available at arXiv:1002.1583.
[30] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Department of Statistical Science
Cornell University
Ithaca, New York 14853
USA
E-mail: pj54@cornell.edu

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: jiashun@stat.cmu.edu